Class-Based Feature Matching across Unrestricted Transformations

Evgeniy Bart and Shimon Ullman

Abstract—We develop a novel method for class-based feature matching across large changes in viewing conditions. The method (called MBE) is based on the property that, when objects share a similar part, the similarity is preserved across viewing conditions. Given a feature and a training set of object images, we first identify the subset of objects that share this feature. The transformation of the feature's appearance across viewing conditions is determined mainly by the properties of the feature, rather than of the object in which it is embedded. Therefore, the transformed feature will be shared by approximately the same set of objects. Based on this consistency requirement, corresponding features can be reliably identified from a set of candidate matches. Unlike previous approaches, the proposed scheme compares feature appearances only in similar viewing conditions, rather than across different viewing conditions. As a result, the scheme is not restricted to locally planar objects or affine transformations. The approach also does not require examples of correct matches. We show that, by using the proposed method, a dense set of accurate correspondences can be obtained. Experimental comparisons demonstrate that matching accuracy is significantly improved over previous schemes. Finally, we show that the scheme can be successfully used for invariant object recognition.

Index Terms—Feature matching, invariant recognition, parts.

1 Introduction

In this paper, we consider the problem of matching corresponding parts of objects across large changes in viewing conditions. The input to the algorithm is a set of images of objects from a given class (such as faces or cars) under different viewing conditions. From this set, the algorithm automatically extracts object parts and matches corresponding parts in different images. An example task is to obtain a gallery of face parts, such as eyes, noses, and mouths, from face images taken under distinctly different viewing angles and illuminations, as illustrated in Fig. 1. These parts can then be used for automatic image interpretation in terms of objects and their parts, invariant recognition, and wide-baseline matching.

The problem of matching object parts is related to the feature-matching problem studied in the past [1], [2], [3], [4]. A basic difference is that, traditionally, the goal of feature matching is to obtain point-to-point correspondences. These point correspondences can then induce correspondences of regions defined by sets of matching points. In contrast, the proposed approach directly finds correspondences between regions that depict significant object parts. If desired, point correspondences can then be recovered from this information. However, for tasks such as object recognition, region correspondence of matching parts is often sufficient [5] (see also Section 5 below). As shown below, region correspondences can be identified more accurately and under more general conditions than previously possible with point correspondences. In addition, region correspondences can be defined under more general conditions than point correspondences. For example, as Fig. 1 illustrates, eye blinking completely changes the eye's appearance. Many points (for example, the pupil) disappear from view, making point correspondences undefined. However, a meaningful correspondence can still be established between the open and the closed eye.

Generic feature-matching methods studied in the past [1], [2], [3], [4] are not restricted to a particular object class and do not require a training set. On the other hand, they make assumptions about objects (such as piecewise planarity) and allowed transformations (such as affinity). Such schemes perform matching reliably across unrestricted affine transformations but may fail when more general transformations are present. In contrast, the scheme proposed here can successfully learn a much more general class of transformations. On the other hand, it requires a training set in order to learn. In addition, it is restricted to those transformations that appear in the training images. For example, with an appropriate training set, the proposed scheme can cope with out-of-plane rotations (Fig. 2), which are distinctly difficult for affine-invariant schemes. On the other hand, since in-plane rotations were not present in this training set, the proposed scheme would fail to match across an in-plane rotation, while an affine-invariant scheme would perform this match automatically. The two approaches are in this sense complementary to each other and should be combined for optimal performance. In this paper, we compare and contrast one approach against the other to illustrate the relative advantages and disadvantages of each.

- *E. Bart is with the California Institute of Technology, 1200 E. California Blvd., MC 136-93, Pasadena, CA 91125, and the Weizmann Institute of Science, Rehovot, Israel. E-mail: eugenenbart@gmail.com.*
- *S. Ullman is with the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Ziskind Building, Room 208, Rehovot, 76100, Israel. E-mail: shimon.ullman@weizmann.ac.il.*

Manuscript received 29 June 2006; revised 13 Feb. 2007; accepted 8 Oct. 2007; published online 16 Oct. 2007.
Recommended for acceptance by L. Van Gool.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0481-0606.
Digital Object Identifier no. 10.1109/TPAMI.2007.70818.

Fig. 1. Illustration of the matching problem. Sample input to the system includes images of the same face in several viewing conditions (top row). The goal is to automatically identify corresponding parts, such as eye (middle row) or mouth (bottom row) regions, in all of these conditions.

The motivation for studying class-based part matching arises for several reasons. First, in several popular schemes of object classification [6], [7], indexing, and retrieval [3], objects are represented by their constituent parts. Theability to identify the same object part under different viewing conditions would enable these schemes to successfully deal with view-invariant recognition. Second, the ability to reliably identify and localize object parts, in addition to the recognition of the entire object, is of interest in its own right. The importance of certain object parts such as the eyes is so high [8] that they can be considered objects of intrinsic interest whose localization can be as important as the localization of the entire object. The ability to identify corresponding object parts under different conditions is therefore an essential aspect of visual recognition. Finally, the problem of feature matching is central to tasks such as wide-baseline stereo and image registration. As discussed in Section 6, the use of class-based information for highly familiar objects such as faces can improve current stereo techniques.

In the scheme described below, class-based information is used to achieve reliable part matching. We use the fact that many objects within a general class (such as faces, cars, airplanes, etc.) share similar parts. Given a part in an image of one object, we identify the set of additional objects of the same class that share this part. The part's transformation across viewing conditions is determined mainly by the properties of the part rather than of the entire object (see the discussion in Section 6). Therefore, the transformed part will be shared by approximately the same set of objects. Since this equivalence requirement is central to the proposed scheme, we call it "matching by equivalence" or MBE. Using equivalence, corresponding parts are identified from a set of candidate matches. This idea is illustrated schematically in Fig. 2. Since the necessity of matching appearances of features across viewing conditions is eliminated, this scheme can identify images of the same object part under much broader conditions than previous schemes (Section 4). Note that if the correct match for the part in question were known for one of the objects, this could be used to infer matches for this part in the remaining objects. However, examples of correct matches are usually unavailable. Our matching method proceeds without requiring such examples and identifies the correct matches in an unsupervised manner.

Fig. 2. Schematic illustration of the proposed matching scheme. First column: A frontal eye part (top) is searched in frontal face images. Faces with similarly looking eyes are marked by white rectangles. Second column: The side view of the same eye is searched in side-view faces. Since the two eye views depict the same object part, they are found in mostly the same faces. Third column: The side view of an unrelated part (mouth) is searched in side-view faces. Since, in general, mouth shapes are not correlated with eyes, this mouth is found in an independent subset of faces. This property can be used to distinguish correct matches (first column with second) from incorrect ones (first column with third).

The remainder of this paper is organized as follows: In the next section, we review previous approaches to the problem of matching local object parts. In Section 3, we describe the proposed MBE scheme. In Section 4, the performance of MBE is evaluated experimentally and compared to several popular algorithms. In Section 5, an application to the problem of view-invariant object recognition is shown. We conclude with general remarks in Section 6.

2 Previous Approaches to Matching

Matching features across views requires predicting how the change of viewing conditions will affect the feature's appearance. In simple cases (for example, when images are related by pure translation), one may assume that the feature's appearance remains constant and only its position changes. This assumption, called "brightness constancy," is used in several well-known feature tracking and optical flow algorithms [1], [9]. Under the brightness constancy assumption, features can be matched using the minimal sum of squared differences criterion.

However, in most practical cases, the variations of viewing conditions affect the feature's appearance considerably. The impact of viewing conditions is often more significant than the impact of the feature's identity [10], [11]. One general approach to cope with this difficulty is to approximate the transformation of the feature's appearance by some parametric model. Typically, affine or low-order polynomial models of illumination and geometry are used (for example, [12]), but more complicated schemes [13], [14] have also been proposed. If estimates of the parameters are known, the features may be matched across transformations (for example, by warping). However, estimating the

parameters is often a difficult task. Most schemes [13], [14] handle it by using video sequences and updating the parameters for each frame. The task of incremental updating is easier because the difference in parameters between successive frames is small. The drawback is the necessity of using video sequences, which are not always available and require additional effort to capture, store, and manipulate. In addition, the approximation provided by common parametric models [12] is only valid for a limited range of transformations.

An approach that does not require an estimation of the transformation parameters is the use of invariant features. Several popular schemes [2], [3], [4], [15] work with so-called affine invariants. The general idea is that the features extracted from an image are normalized with respect to affine transformations. Therefore, features differing by an affine transformation have identical representations and can be matched directly, without the need to estimate the parameters of the transformation. A drawback of this approach is that affine approximation holds only for rigid transformations of locally planar scenes and affine illumination changes. (This is less restrictive than it may appear. For example, smoothly deforming objects such as a bending magazine can be approximated by locally planar regions [16].) However, many natural objects (such as faces) are not planar and are subject to nonrigid transformations such as facial expression and therefore will be difficult to model in this framework. In addition, illumination changes perturb image intensity in a highly nonlinear manner due to such factors as specularities and cast shadows. Finally, affine transformations cannot model changes of appearance resulting from physical deformations of the object (such as eye blinking, articulation, and lip movements during speech). When affine approximation does not hold, affine invariants by themselves will be insufficient to achieve good performance (see the experiments in Section 4). Additional processing might be needed in such cases to perform matching, as in [16]. Invariants more general than affine exist [17] but are usually sparse and, therefore, insufficient for most tasks. An additional drawback of methods utilizing invariants is the lack of control over the features: Since not all image points are invariant, it is impossible to match a particular point of interest; only the points identified as invariant can be used. A scheme that can cope with complex intensity transformations was described in [18]. However, this technique can only handle affine geometric distortions and requires a 3D model for more complex transformations.

Recently, a method based on affine invariants [16] has been used to achieve view-invariant object recognition. The method (called "Image Exploration") extends affine invariants to cope with clutter, occlusion, and nonrigid deformations. In this paper, we explore a different complementary approach to matching, based on an equivalence constraint.

Correspondences between object parts have been used for alignment and matching [5]. However, the correspondences were specified manually [5]. In [19], matching regions were learned from video sequences. A similar idea was described in [20], where the shape of each object part was averaged over the transformations. However, these methods require that examples of correct matches be provided, in contrast to the MBE scheme proposed here.

3 Class-Based Matching by Fragment Equivalence

In this section, we describe the proposed MBE algorithm for class-based matching. In Section 3.1, the idea of utilizing the equivalence criterion for matching individual object parts is presented. In Section 3.2, we describe how the accuracy of matching is improved by exploiting geometric constraints. The final algorithm, which combines the appearance and geometry, is described in Section 3.3.

3.1 Matching Object Parts

Before describing the part-matching method, we first briefly describe the relevant aspects of fragment-based object representation [6], [7], [21] that are used by the current method. In this scheme, objects from a general class (such as cars or faces) are represented by their constituent parts. For example, parts for face images typically include different types of eyes, mouths, etc. Image patches, or *fragments*, are used to depict the appearance of each object part. Fragments are extracted automatically from example images in the learning stage. Each fragment is searched for in the images using normalized correlation or another suitable similarity measure such as that in [22]. As described below, MBE uses this measure only to compare fragments in similar viewing conditions. For this reason, the choice of similarity measure is not crucial and even a simple measure such as normalized correlation gives good results. In our experiments, we used the absolute value of normalized cross correlation (NCC), given by

$$NCC(p,F) = \frac{\frac{1}{N}\sum_{x,y}(p(x,y)-\overline{p})(F(x,y)-\overline{F})}{\sigma_p \sigma_F}. \quad (1)$$

Here, $F(x,y)$ is the fragment, $p(x,y)$ is an image patch of the same size as F, N is the number of pixels in the fragment, $\overline{p}$ and $\overline{F}$ are the means, and σ_p and σ_F are the standard deviations of the intensities of p and F. Image patches at all candidate locations are compared with F and the location with the highest correlation is selected. This highest correlation value is called the *activation* of the fragment in the image. Since the absolute value of the normalized correlation is taken, this activation is a continuous value in [0, 1]. When the activation exceeds a predetermined threshold, the fragment is considered present, or *active*, in the image. An object is represented by the set of fragments that are active in it.

The simplest way to utilize this representation for part matching between two images of the same object is to directly match a fragment from one image with the other image by normalized correlation. However, this method performs poorly when significant variations in viewing conditions are present [10]. To obtain a reliable match between the same object part under different viewing conditions, consider two fragments, F and F', depicting the same object part P in different viewing conditions C and C'. We rely on the fact that part P itself does not change during the transformation, although its appearance changes from F to F'. Therefore, fragment F, used with images taken under conditions C, plays an equivalent role to that played by F' in conditions C'. For example, if F is active in the image of some object under conditions C, then F' will be active in the image of the same object under conditions C'. In other words, F and F' will be consistently

detected in images of the same objects, F in conditions C and F' in conditions C'.

Given an arbitrary pair F, F' of fragments, this consistency can be used to test whether F matches F'. For this test, a set of images of additional objects in viewing conditions C and C' is used. To simplify the presentation, we assume here that the viewing conditions are known, that is, that, for each image, it is known whether it was taken under C or C'. This assumption will be removed in Section 4.1. We also assume that the viewing conditions are the same for all objects. This assumption will be relaxed in Section 4.2. The subset S of objects in viewing conditions C in which F is active is identified using normalized correlation. This task is straightforward because the viewing conditions of the fragment and the images are similar. Similarly, the subset S' of objects in viewing conditions C' in which F' is active is identified. As discussed above, the presence of one fragment reliably predicts the presence of its matching fragment. Therefore, the sets S and S' will be similar. Conversely, for nonmatching fragments, these sets will, in general, be significantly less similar. This is because nonmatching fragments represent different object parts. In general, different object parts are not highly correlated in different images. Therefore, the presence of one fragment will be independent of the presence of the other. This matching by consistency is illustrated schematically in Fig. 2.

Next, we describe how the consistency of two fragments F and F' is evaluated in practice. The main difference from the schematic description above is that continuous unthresholded activations are used, rather than presence or absence (which is based on thresholded values). The reason is that considering only the presence or absence of a fragment in an image discards important information about the strength of activation.

Assume that a set of images $I_1, \ldots, I_n$ of n objects taken under conditions C and a set of images $I'_1, \ldots, I'_n$ of the same objects taken under conditions C' are given. (These sets will be called "validation database.") For the fragment F, let A_k be the activation of F in I_k. Note that A_k is a continuous unthresholded value in [0, 1]. These values are combined into an n-dimensional activation vector A. This vector A can be thought of as a descriptor of the fragment F. Similarly, we calculate the activation vector A' for F' by setting A'_k to the activation of F' in I'_k.

The task is now to evaluate the similarity of the two descriptors A, A'. To achieve this, we recall the equivalence constraint. First, note that A defines a ranking of images by the activation of fragment F. For example, the largest entry in A corresponds to the image in which F was most active. Similarly, A' defines a ranking of images by the activation of F'. Matching fragments play equivalent roles in their respective viewing conditions and therefore will define similar rankings of the corresponding images. The similarity of rankings can be measured by Kendall's τ rank correlation coefficient [23]. This coefficient is $+1$ when the rankings induced by A and A' agree perfectly, is -1 when the rankings are opposite of each other and has intermediate values for intermediate degrees of agreement between rankings. Since it is convenient to have a consistency range from 0 to 1, the value

$$C(F, F') = \frac{\tau(A, A') + 1}{2} \tag{2}$$

was used. $C(F, F')$ ranges from 0 to 1, with 1 indicating perfect consistency and 0 indicating complete inconsistency. Experiments were also performed using Spearman's rank correlation coefficient, as well as similarity measures of binary (thresholded) activation values. However, these measures performed more poorly than Kendall's τ.

The class-based matching algorithm can now be described as follows: Given two images, I (called "source image") and I' (called "target image"), of the same object taken in conditions C and C', the task is to find a dense set of correspondences between I and I'. For every location in I, consider a small image patch F (called "source fragment") at that location that depicts some object part P under C. In order to find the matching fragment F', consider all candidate target fragments in I' and select the most consistent fragment F' as the match. Examples of matches obtained by this algorithm are shown in Fig. 3.

3.2 Using Fragment Pyramids to Improve Accuracy

The previous section explained how fragment consistency is used to evaluate the likelihood of an individual match. However, simply matching the two most consistent features is not the optimal strategy because factors such as clutter and within-object redundancy may cause matching errors (Fig. 4). A common strategy is to employ some geometric constraints between features to improve the matching accuracy of individual features.

To derive geometric constraints, several existing schemes assume some parametric model, such as a homography, of the global scene transformation [2], [3], [15]. However, such an assumption is often too restrictive in practice and, therefore, more general geometric constraints are desirable. The constraint incorporated in the equivalence-based matching scheme is a simple proximity assumption. Intuitively, we assume that if two object parts are located close to each other in one image, they are likely to remain close in other images (see the discussion in Section 6).

To impose the proximity constraint, we use a hierarchical representation of the proximity relations between object parts. A high-level overview of this hierarchical representation is given in this section. The description of the details is deferred to Section 3.3, where the complete matching scheme is presented.

To represent the proximity relations between object parts by a hierarchical structure, the image is covered by progressively larger fragments until a single fragment covers the entire image. A smaller fragment whose central point is inside the area covered by a larger fragment is considered a child of the larger fragment in the hierarchy. To make the inference efficient, a fragment is allowed to have only one parent. To ensure this condition, the large fragments are selected to *tile* the image, that is, to cover the entire image area and be nonoverlapping. Since, by construction, the lowest level fragments have no children, the tiling requirement is not necessary at this level. Therefore, fragments on the lowest level are created at every image location and are allowed to overlap.

In this hierarchy, any two fragments will eventually have a common ancestor. The more spatially close the two

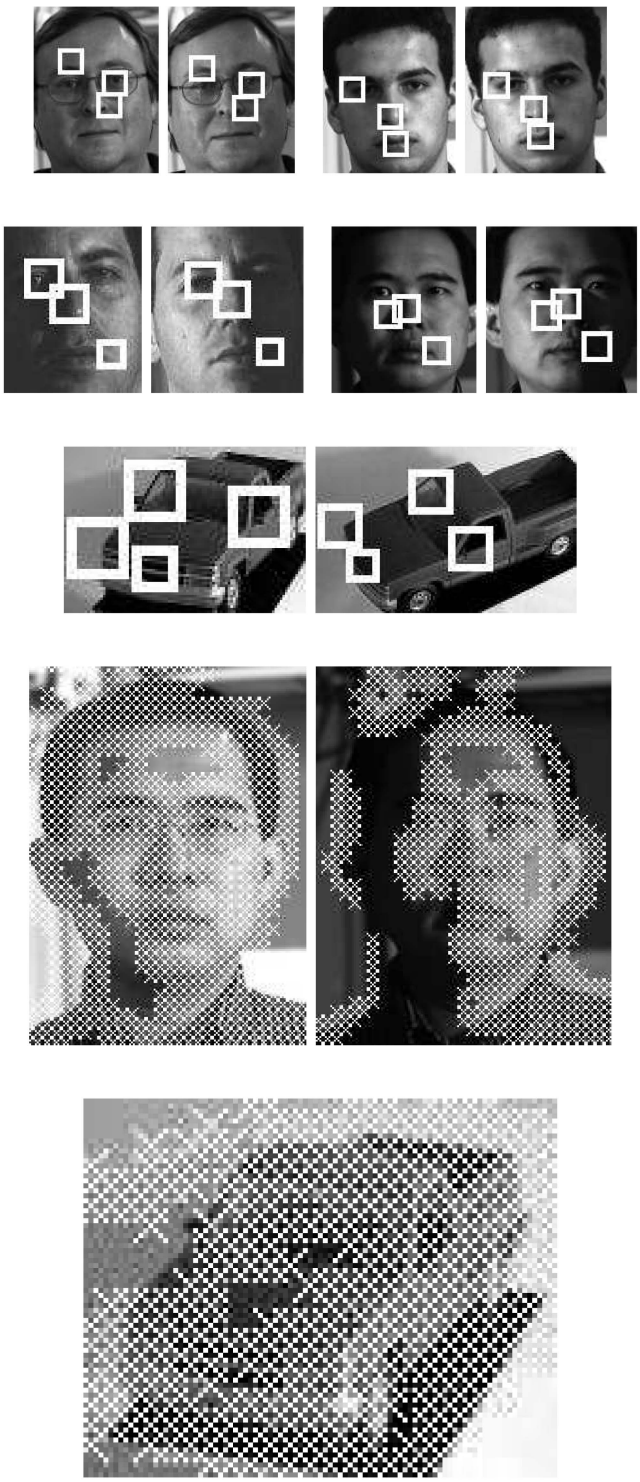


Fig. 3. Examples of matches obtained automatically by the algorithm. (First row) Easy illumination data set. (Second row) Hard illumination data set. (Third row) Cars data set. (See Sections 4 and 5.) The squares mark the positions of corresponding fragments. Only a few matches are shown, the total number of matches was several hundred for each pair of images. Note that object parts are matched accurately despite significant cast shadows and pose changes. (Fourth row and fifth row) Density of matches. Locations at which features were matched are marked by white crosses. (Fourth row, left) Easy illumination data set. (Fourth row, right) Hard illumination data set. (Fifth row) Cars data set. Note that matches cover densely the object of interest.

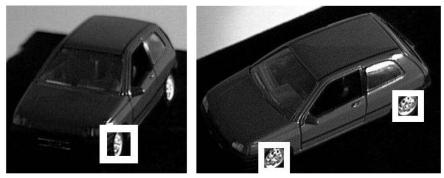


Fig. 4. Due to within-object redundancy, two equally good matches are possible. To disambiguate such cases, geometric constraints are introduced, as explained in Section 3.2.

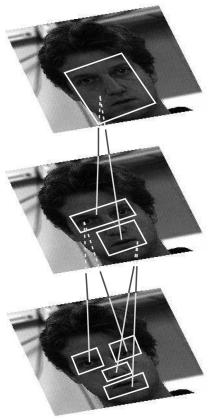


Fig. 5. An object is covered by fragments of progressively larger size, depicted by white rectangles. Smaller fragments that fall inside larger ones are considered to be their children, as indicated by the gray lines. The resulting hierarchical structure is called the fragment pyramid. Note that, in this schematic illustration, high-level fragments do not tile the image.

fragments are, the lower the level of this common ancestor will be. In particular, two neighboring fragments are likely to have an immediate common parent. In contrast, two distant fragments are likely to be more separated in the hierarchy. In this manner, proximity will be represented by the hierarchy, as illustrated in Fig. 5. The resulting structure is called the fragment pyramid.

Note that two fragments that are close to each other in the image but which happen to fall on the opposite sides of a higher level tile will only have a distant common ancestor. This is a potential disadvantage of using the fragment pyramid to represent proximity. However, in practice, this does not seem to be an issue (Section 4).

Since proximity relations are roughly preserved across the transformation of viewing conditions, the hierarchical structure in the source and target images should be similar. Deviations from the common structure generally indicate unlikely matches. Occasional exceptions to this general rule can be tolerated since probabilistic (rather than hard) constraints are used in the complete scheme. This tolerance is sufficient to allow accurate matching even when the proximity relations are preserved only approximately (see the experiments in Section 4 and the additional discussion in Section 6).

This hierarchical structure is flexible enough to represent complex nonplanar transformations. In addition, the tree representation allows efficient inference: It avoids iterative computations and identifies the optimal match with two passes over the hierarchy. The next section describes how the fragment pyramid is used to impose the proximity constraint and how it is combined with the consistency measure to produce the final matching.

Input: source image I, target image I'
Output: pairs of fragments f, f', such that f in I matches f' in I'

1) **Build the source pyramid.**
   a) **Select the leaves of the pyramid.** The leaves are the fragments in I for which matches are to be found. Typically, the user supplies a set of locations and fragment sizes to be matched. This set can be arbitrary. If no pre-specified set is available, locations on a regular grid are selected automatically. At each location, a fragment is created.
   b) **Complete the source pyramid.** Given the size of fragments on the previously created level, increase this size by a factor of $s = 2$. Tile the image by a regular grid of fragments of the current size. Repeat this step until a single fragment covers the entire image.
2) **Construct the candidate target fragments.** Create candidate target fragments at every location in the target image. Pyramid requirements are not enforced in the target image.
3) **Find the optimal matches.** Apply the two-pass inference algorithm (Figure 7) to maximize the probability of matches given by eq. (4). The values of the factors in the decomposition are calculated from eqs. (7), (9).

Fig. 6. MBE algorithm. For details, see Section 3.3.

3.3 Combining Consistency and Proximity Constraints

The final matching strategy combines the two factors described so far, the likelihood of individual matches (measured by fragment consistency) and the similarity of geometrical structure (measured by the fragment pyramid). In Fig. 6, an overview of the algorithm is presented. Below, different stages of the algorithm are described in detail.

First, a fragment pyramid is constructed from the source image. The process is started by creating fragments of certain initial size at every image location. The size of the first-level fragments is then increased by a fixed factor s to determine the size of the second-level fragments. In our experiments, we used $s = 2$. (The scheme was not sensitive to the exact value of s.) Larger fragments that tile the source image are created and parent-child relationships are determined. The process of increasing the fragment size is repeated until a single fragment covers the entire image. This fragment is considered the root node of the hierarchy.

The choice of the initial fragment size is an important parameter of the scheme and may affect significantly the accuracy of matching. This is illustrated by the following examples: Consider the extreme case of two fragments F and F' of size 1 pixel each. Since a single-pixel fragment can be detected in any image, every element of the activation vectors of both F and F' will be equal to 1. The two fragments will then be entirely consistent, regardless of whether they actually originate at matching locations. In the other extreme, a fragment that is too large (for example, 50 percent of the image area) will be highly specific to the image from which it was extracted and will be rarely detected in other images. The two large fragments will therefore have activation vectors of almost all zeros and may again be highly consistent regardless of whether they actually match. To avoid this problem, fragment size should be selected so that the fragments represent meaningful object parts. This size is currently set manually to roughly 10-20 percent of the image size. We have also tried to select the optimal size automatically by maximizing $I(F; ID)$, the mutual information between fragment appearance and object identity, as suggested in [7]. Initial experiments with this scheme gave promising results; however, in most of the experiments reported below, the size was set manually.

An additional consideration for setting the fragment size is the following: The proposed algorithm produces region correspondences. Point correspondences are extracted by matching the centers of corresponding regions. Conceivably, the larger the matched regions are, the more uncertain point matches might be, even if region correspondences were accurate. However, the experiments below show that MBE attains useful accuracy even with large fragment sizes (for example, an average error of 5 pixels even when the fragments were of size 51×51 pixels, as shown in Section 4).

The number of levels in the pyramid also depends on the size of the first-level fragments. If the first-level fragments are large, the pyramid will have too few levels and will only coarsely represent spatial relations between different fragments. This could potentially limit the accuracy of the scheme. However, the experiments in Section 4 demonstrate that, in practice, accurate matches are obtained even with fragments of size 51×51 pixels (roughly $1/5$ of the image size).

After building the fragment pyramid in the source image, candidate target fragments are created in the target image. Note that the fragment pyramid is constructed for the source image only. Candidate target fragments are created at every position in the target image and the constraints of tiling are not enforced. Instead, for each level of the source fragment pyramid, candidate target fragments at all positions in the target image are created and considered as possible matches.

The size of the candidate target fragments is also an important consideration. If the source and target images are of the same scale, then it is natural to consider target fragments of the same size as source fragments at the corresponding pyramid level. If the target image is scaled by a factor S, then it is natural to scale the target fragments by the same factor relative to the source fragments. In most cases, however, the scale is not known. In this situation, candidate target fragments at multiple scales can be created that cover the entire range of possible scales. For example, if it is believed that the target image is scaled by a factor S between 0.5 and 2, then candidate target fragments within this range of scales need to be created. Note that, in this case, multiple candidate matches will be available instead of one at each location in the target image. For each source

fragment, the algorithm will then select both the location and the scale of the most consistent target fragment.

The problem we now face is to establish matches between fragments in the source and target images. We first describe the general strategy (3)-(6) and, then, the simplified approximation used in the implementation (7)-(9) to incorporate the spatial constraints. Let $\mathcal{F}$ denote the set of all fragments in the pyramid. Denote by X the unknown vector of matches. This vector has length $|\mathcal{F}|$ (the number of elements in $\mathcal{F}$) and $X_f = f'$ if fragment f matches f' (X_f is the entry corresponding to fragment f). The unprimed variables below refer to source fragments and the primed variables refer to target fragments. Denote by Y the set of observations, which, in our case, include the consistency values $C(f, f')$ for every pair of fragments f, f' at the same pyramid level, with f taken from the source image and f' taken from the target image. The commonly used maximum a posteriori (MAP) estimate of the vector of matches is given by

$$\hat{X} = \arg\max_X P(X, Y), \tag{3}$$

where $P(X, Y)$ is the joint probability distribution of X and Y. To use this estimate, it is desirable to simplify $P(X, Y)$. This simplification is performed by making a number of standard and plausible independence assumptions.

First, we assume that the fragments depend only on their parents in the pyramid and are independent given the parent's correct match. We also assume that the observations $C(f, g')$ involving some source fragment f are determined by the value of X_f (its true match) and g' and are therefore independent of all other variables given X_f. These two assumptions are standard for similar models [24].

Combining the assumptions above results in the following factorization:

$$P(X, Y) = P(X_R) \prod_{f \in \mathcal{F} \setminus \{R\}} P(X_f | X_{\pi(f)}) \prod_{f \in \mathcal{F}} P(Y_f | X_f). \tag{4}$$

Here, R is the root fragment, $\pi(f)$ is the parent of f in the pyramid, and Y_f is the set of all observed consistency values that involve the fragment f:

$$Y_f = \{C(f, g') : g' \text{ at the same level as } f\}. \tag{5}$$

Moreover, $P(Y_f | X_f)$ can be further simplified to

$$P(Y_f | X_f = f') = \prod_{g' \in M'(f)} P(C(f, g') | X_f = f'), \tag{6}$$

where $M'(f)$ is the set of the possible matches of f, that is, $M'(f)$ is the set of candidate target fragments g' at the same level as f. See the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70818, for a more detailed justification of this decomposition. Under this decomposition, the MAP estimate of each fragment's best match can be efficiently calculated using the standard Viterbi-like inference algorithm [24], [25] summarized in Fig. 7. This algorithm is guaranteed to converge in two passes and produce the optimal (in the MAP sense) solution.

Bottom-up pass. In every node f, starting at the leaves, calculate functions $Q_f(f', f'_\pi)$, $P_f(f'_\pi), F^*_f(f'_\pi)$ for all values of f', f'_π, as follows:
For leaves:

$$Q_f(f', f'_\pi) = P(Y_f | X_f = f') P(X_f = f' | X_{\pi(f)} = f'_\pi)$$
$$P_f(f'_\pi) = \max_{f'} Q_f(f', f'_\pi)$$
$$F^*_f(f'_\pi) = \arg\max_{f'} Q_f(f', f'_\pi)$$

For all non-leaves except the root:

$$Q_f(f', f'_\pi) = P(Y_f | X_f = f') P(X_f = f' | X_{\pi(f)} = f'_\pi) \cdot \cdot \prod_{t \in \mathcal{C}(f)} P_t(f')$$
$$P_f(f'_\pi) = \max_{f'} Q_f(f', f'_\pi)$$
$$F^*_f(f'_\pi) = \arg\max_{f'} Q_f(f', f'_\pi)$$

Here $\mathcal{C}(f)$ denotes the set of children of node f. The index f' ranges over $M'(f)$, all candidate matches of the current node. The index f'_π ranges over $M'(\pi(f))$, all candidate matches of the parent of the current node. This parent is determined based on the fragments pyramid built from the source image.

Top-down pass. For every node f, starting at the root, recover $\hat{f}'_f$, the MAP estimate of the match, as follows:
For the root:

$$\hat{R}'_R = \arg\max_{f'} P(Y_R | X_R = f') P(X_R = f') \prod_{t \in \mathcal{C}(f)} P_t(f')$$

For all nodes except the root:

$$\hat{f}'_f = F^*_f(\hat{f}'_{\pi(f)})$$

Fig. 7. Two-pass MAP estimation algorithm. For the derivation and proof, see [24].

Next, we describe how the individual factors in (4) are estimated. We used approximations that simplified the model construction and proved sufficient in practice for using the proximity constraint. First, consider $P(C(f, g') | X_f = f')$. Intuitively, if $X_f = f'$ (that is, f matches f'), the similarity value $C(f, f')$ is expected to be high. In contrast, there is, in general, no reason for $C(f, g')$ ($g' \neq f'$) to be high; therefore, small values of $C(f, g')$ have a much higher probability than small values of $C(f, f')$. Therefore, the distributions $P(C(f, f') | X_f = f')$ and $P(C(f, g') | X_f = f')$ for $g' \neq f'$ are qualitatively different. On the other hand, $P(C(f, f') | X_f = f')$ for different fragments f and f' are qualitatively similar. For simplicity, we assume that the distribution $P(C(f, f') | X_f = f')$ is the same for all fragments f and f'. For the same reason, we assume that the distribution $P(C(f, g') | X_f = f')$ is the same for all f, f', and $g' \neq f'$. Therefore, only these two distributions need to be evaluated.

In principle, both distributions could be estimated from training data, but this would require examples of correct matches to be available during training. To avoid this

requirement, we approximate $P(C(f,f')|X_f=f')$ and $P(C(f,g')|X_f=f')$ by distributions that satisfy some intuitively appealing assumptions. These approximations performed well in our experiments and the algorithm was insensitive to the precise form of the approximation.

First, we estimate $P(C(f,f')|X_f=f')$. Intuitively, if f matches f', then high values of $C(f,f')$ are more likely than low values. The estimate

$$P(C(f,f')|X_f=f')=\frac{1}{Z_1}[C(f,f')]^{\gamma} \tag{7}$$

(where Z_1 is an appropriate normalization factor) qualitatively agrees with this intuition and was used in the experiments. The value $\gamma=8$ performed best in our experiments and all of the experiments below used this setting of γ. The algorithm was not very sensitive to the precise value of γ, but it was important to have $\gamma>1$ to make the maxima of $C(f,f')$ sharper. For a more detailed justification of (7), see the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70818.

We assumed a uniform distribution for $P(C(f,g')|X_f=f')$. The intuition here is that two nonmatching fragments f and g' are independent; therefore, all consistency values are equally likely. In particular, high consistency for a nonmatching pair can be obtained by chance. In this case, $P(Y_f|X_f=f')$ can be written simply as

$$P(Y_f|X_f=f')=\frac{1}{Z_1}[C(f,f')]^{\gamma}, \tag{8}$$

where Z_1 is adjusted so that $P(Y_f|X_f=f')$ is a valid probability distribution. This approximation was used in the final scheme. For more details, see the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70818.

The distribution $P(X_f=f'|X_{\pi(f)}=f'_{\pi})$ implements the proximity constraint, using the hierarchical structure. Assume that $\pi(f)$ matches f'_{π}. By definition, f is a child of $\pi(f)$. The similarity of the hierarchical structure requires f' to be a child of f'_{π}. Therefore, $P(X_f=f'|X_{\pi(f)}=f'_{\pi})$ should be high if the smaller fragment f' falls inside the larger fragment f'_{π} and should decrease when f' becomes more distant from f'_{π}. The following estimate conforming to the qualitative requirements listed above was used:

$$P(X_f=f'|X_{\pi(f)}=f'_{\pi})=\frac{1}{Z_2}\frac{1}{d(f',f'_{\pi})^{\beta}}. \tag{9}$$

Here, $d(f',f'_{\pi})$ is the distance from the center of f' to the center of f'_{π} and Z_2 is the appropriate normalization factor, calculated so that the right-hand side of (9) sums to 1. In our experiments, this estimate was sufficient to obtain good performance. We have used $\beta=0.25$ to make the maxima of d broader. The algorithm was not sensitive to the exact value. For more details, see the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70818.

Finally, the uniform distribution was used for the prior probability of the root node match $P(X_R=R')$.

Bottom-up pass. In every node f, starting at the leaves, calculate functions $Q_f(f',f'_{\pi})$, $P_f(f'_{\pi})$, $F^*_f(f'_{\pi})$ for all values of f', f'_{π}, as follows:
For leaves:

$$Q_f(f',f'_{\pi})=\frac{1}{Z_1}[C(f,f')]^8\frac{1}{Z_2}\frac{1}{d(f',f'_{\pi})^{0.25}}$$

$$P_f(f'_{\pi})=\max_{f'}Q_f(f',f'_{\pi})$$

$$F^*_f(f'_{\pi})=\arg\max_{f'}Q_f(f',f'_{\pi})$$

For all non-leaves except the root:

$$Q_f(f',f'_{\pi}) = \frac{1}{Z_1}[C(f,f')]^8\frac{1}{Z_2}\frac{1}{d(f',f'_{\pi})^{0.25}}\cdot \cdot\prod_{t\in\mathcal{C}(f)}P_t(f')$$

$$P_f(f'_{\pi})=\max_{f'}Q_f(f',f'_{\pi})$$

$$F^*_f(f'_{\pi})=\arg\max_{f'}Q_f(f',f'_{\pi})$$

Top-down pass. For every node f, starting at the root, recover $\hat{f}'_f$, the MAP estimate of the match, as follows:
For the root:

$$\hat{R}'_R=\arg\max_{f'}[C(R,f')]^8\prod_{t\in\mathcal{C}(f)}P_t(f')$$

For all nodes except the root:

$$\hat{f}'_f = F^*_f(\hat{f}'_{\pi(f)})$$

Fig. 8. The final matching algorithm. For notation, see Fig. 7 and Section 3.3.

The final matching algorithm is obtained by substituting the estimates above into the algorithm in Fig. 7. This final algorithm is summarized in Fig. 8. This simplified version has the advantage of making the model learning straightforward. In particular, no parameters need to be estimated from training data. Only the consistency values $C(f,f')$ and the distances $d(f',f'_{\pi})$ are calculated. The normalization factors in (7), (8), and (9) are calculated from these values in a straightforward manner. Since the proposed model was not sensitive to these simplifications, the resulting algorithm still gives useful results.

4 Experimental Evaluation

4.1 Illumination and Pose

In this section, we compare the accuracy of the proposed fragment equivalence scheme to several well-known matching schemes, namely, KLT [1], Black's robust optical flow [9], and affine-invariant features [3]. (See Section 2 for the description of these schemes.) The KLT implementation available at [26], Black's original implementation of robust optical flow available at [27], and Mikolajczyk's original implementation of invariant features available at [28] were used for the experiments.

Note that KLT and robust optical flow require the matched images to be similar. This is usually achieved by using

TABLE 1
Average Errors in Matches $\pm$ Standard Deviation, in Pixels

Algorithm	Easy illumination	Hard Illumination	Pose
Affine invariants	82 ± 71	105 ± 69	56 ± 34
KLT	58 ± 43	74 ± 34	89 ± 29
Robust optical flow	73 ± 5	74 ± 5	125 ± 20
Equivalence, no pyramid	7 ± 14	21 ± 27	24 ± 19
Equivalence, with pyramid	5 ± 8	13 ± 15	17 ± 11

adjacent frames of a video sequence. Videos are frequently unavailable; in particular, only still images were used in the experiments below. As expected, this significantly reduces the performance of KLT and robust optical flow.

KLT and robust optical flow were applied to each image pair independently. For the affine invariants scheme, invariant points for each image were calculated. Region matching was then performed by selecting, for each source point, the target point with the most similar descriptor. The similarity of the descriptors was measured by the Mahalanobis distance, using the covariance matrix estimated during the training stage from all of the images in the database (a separate matrix was used for each experiment below and each database). This evaluation is identical to the published description of the algorithm [3].

The current fragment equivalence scheme was evaluated by using as the validation database the entire image set excluding the two images being matched. The size of first-level source fragments was set to 51×51 pixels. Candidate target fragments were of the same size as source fragments.

Recall that, when describing MBE in Section 3.1, we have made an assumption that it is known for each image in the validation data set whether its viewing conditions correspond to the source or target image. In practice, images taken under different conditions may be mixed and it is convenient to relax this assumption. A simple strategy is to detect each fragment in every image of the given object and select the highest activation. We have compared this strategy with detecting a fragment only in images under known correct viewing conditions. The results showed that restricting the detection to the correct viewing conditions does not yield a significant increase in performance. The explanation is that, due to the fact that viewing conditions significantly change the appearance of object parts, fragments are automatically detected almost exclusively in the correct viewing conditions. The conclusion is that, when images are not labeled by viewing conditions, fragments can still be extracted correctly by detecting them in all images, without compromising performance. All of the results presented here have been obtained using this strategy.

The accuracy of the proposed method with and without using fragment pyramids was also compared. This comparison demonstrated that the fragment pyramid significantly improves the matching accuracy (Table 1). This comparison also shows that, even without using pyramids, class-based knowledge significantly improves matching over the generic matching schemes.

The data sets used for each of the experiments are described below. The performance of the evaluated algorithms is summarized in Table 1. Errors are expressed as the distance (in pixels) between the correct match and the match returned by the algorithm. A lower error is better. In the illumination data sets, the images of the same object under different illuminations are produced aligned. In the pose data set, the ground truth was determined manually.

4.1.1 Description of the Data Sets

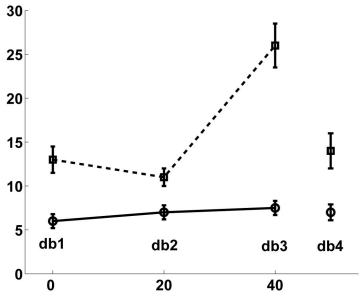
1. *Illumination data set—easy*. Frontal face images from the PIE database [29] were used in this experiment. The data set contains images of 68 individuals. The images were taken with normal (ceiling) room illumination. In addition, in the source images, a light flash from the far right direction (approximately 90 degrees) was added and, in the target images, a flash from the far left (approximately 90 degrees) was added. The original 486×640 images were low-pass filtered and downsampled to size 243×320 pixels. Examples are shown in Fig. 11.
2. *Illumination data set—hard*. Images were again taken from the PIE database. They were similar to the previous set but with the room illumination turned off. As shown in Fig. 11, this makes the changes introduced by illumination variations significantly more severe. In particular, illumination can no longer be approximated by a local affine transformation of intensities.
3. *Pose data set*. A subset of 100 face images from the FERET database [30] was used in this experiment. Frontal images were used as the source and half-profile images (images facing +40 degrees to the right) were used as the target. The images were of size 384×256 pixels.

4.1.2 Summary of the Results

The matching accuracy of the evaluated algorithms is summarized in Table 1. As can be seen, algorithms that rely on brightness constancy (KLT and optical flow) perform relatively poorly in all tasks. This is due to significant changes in appearance between the source and target images.

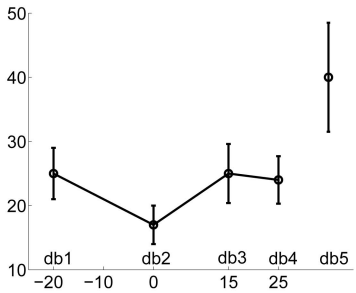
Affine invariants perform reasonably well in the easy illumination task. Although most of the matches are incorrect (average error of 82 pixels, as shown in Table 1), 27 percent of the matches are quite accurate, with an error of less than 6.4 pixels, which is 2 percent of the image size

	flash 1	flash 2
db 1	$-40°$	$+40°$
db 2	$-40°$	$+20°$
db 3	$-40°$	$0°$
db 4	$-20°$	$+20°$



(a)

	face 1	face 2
db 1	$0°$	$+60°$
db 2	$0°$	$+40°$
db 3	$0°$	$+25°$
db 4	$0°$	$+15°$
db 5	$+15°$	$+25°$



(b)

Fig. 9. Matching accuracy as a function of the viewing conditions in the validation data set. (a) Illumination data sets. Left: Flash directions. Right: Matching accuracy. The markings db1-db4 correspond to databases 1-4. Y: Average matching error (a lower value is better). Bars: Atandard error of the mean. For databases 1-3, the X axis is the angle of deviation of the validation database from the testing images (larger values signify larger deviation). For database 4, the X axis location is arbitrary. Solid line, circles: Room lights on. Dashed line, squares: Room lights off. (b) Pose data sets. Left: Poses. Right: Matching accuracy. The markings db1-db5 correspond to databases 1-5. For databases 1-4, the X axis is the angle of deviation of the validation database from the testing images. For database 5, the X axis location is arbitrary.

(data not shown). Correct matches could therefore be identified using RANSAC or a similar procedure [31]. However, in the more difficult illumination task and in the pose task, the performance of affine invariants significantly deteriorates. This is due to the fact that appearance changes induced by cast shadows and pose change cannot be approximated accurately by local affine models of illumination and geometry.

The proposed fragment equivalence scheme performs significantly better than the alternative schemes in all tasks. In the easy illumination task, 73 percent of the matches are accurate (with an error of less than 6.4 pixels, 2 percent of the image size). As shown in Section 5, this accuracy is sufficient for tasks such as object recognition. Some examples of matches are shown in Fig. 3. Fig. 3e shows that a dense set of matches is obtained. The matching performance remains high in all tasks, although the hard illumination task and the pose task are more difficult for this scheme as well. The results demonstrate that matching across significant changes in viewing conditions can be achieved by using appropriate class-based information.

4.2 Sensitivity to Viewing Conditions

In the experiments described above, the viewing conditions in the validation database exactly matched the viewing conditions of the source and target images. This degree of similarity of viewing conditions may be difficult to achieve in practical applications. Below, we describe several experiments in which the viewing conditions in the validation database deviate systematically from the source/target viewing conditions.

4.2.1 Illumination with Room Lights

Frontal face images from the PIE database [29] were used in this experiment. The images were similar to the "easy illumination data set" (Section 4.1.1), except for flash directions. Normal (ceiling) room illumination was on for all images. In addition, a flash was added for each image. In the source images, the flash direction was always $-40°$ (negative value indicates direction to the left). In the target images, the flash direction was always $+40°$ (positive value indicates direction to the right). Four different validation databases were constructed. Each contained two images per person and the two images had different flash directions. These directions are summarized in Fig. 9. In validation database 1, the viewing conditions exactly match the source/target conditions. This database provides the base for comparison. For validation databases 2 and 3, the illumination for the first image matches the illumination of the source image, while the illumination of the second image varies systematically from the target illumination. In validation database 4, the illumination directions for both images are different from the source/target illumination. Note that, even though the validation database changes, the viewing conditions for the source and target images remain the same. This allows direct comparisons of matching accuracy across different validation databases. (If, instead, a single validation database was used and the viewing conditions of the source/target images were varied, the difficulty of the matching task would change, making direct comparisons impossible.) The matching accuracy is plotted in Fig. 9. As can be seen, the decrease in performance is insignificant compared to ideal conditions.

4.2.2 Illumination without Room Lights

The setup for this experiment was similar to that in Section 4.2.1, except that images were acquired with the room illumination turned off. The matching accuracy is plotted in Fig. 9. As can be seen, the performance remains stable for $20°$ of illumination direction change. When illumination changes by $40°$, the performance deteriorates significantly. This increased sensitivity to viewing conditions (compared to that in Section 4.2.1) is due to the fact that the illumination direction now affects the image much more significantly.

4.2.3 Sensitivity to Pose

The setup was similar to that in Section 4.2.1, but face orientation was varied instead of illumination direction. Source images were always frontal (0 degrees) and target images were always half-profiles. Five different validation databases were constructed; the orientations are summarized in Fig. 9. The accuracy is plotted in Fig. 9. As can be seen, the decrease in performance is modest for small

Fig. 10. Examples of matches on the buildings data set (Section 4.4).

rotations. However, the scheme is much more sensitive to the condition when the validation database matches neither the source nor the target image.

The conclusion from the experiments in this section is that the conditions in the validation database need not exactly match the conditions of the source and target images.

4.3 Scale Changes

Scale changes are prevalent in uncontrolled image collections. In this section, we evaluate the ability of MBE to perform matching across scale changes. To allow controlled and careful evaluation of performance as a function of scale, we artificially changed the image resolution to obtain scale changes. Some examples of matches across scale in real images are shown in Section 4.4 (Fig. 10).

4.3.1 Naive Scheme

Again, we have used the PIE database [29] for experiments. To remove all sources of image variability except for scale, we have used frontal face images taken with neutral illumination. The original images in the database have a resolution of 486×640 pixels. The source images were obtained by rescaling the original image to 243×320 pixels. The target images were obtained by rescaling the same original image to a certain scale S that varied from experiment to experiment. For scale S, the size of the target images was $486 \cdot S/2 \times 640 \cdot S/2$ pixels. For example, for $S = 1$, the size was 243×320 pixels (that is, the source and target images were identical), for $S = 2$, the target image was twice as large as the source image, and, for $S = 0.5$, the target image was half the size of the source image. Within a single experiment, all target images had the same scale S. Experiments with $S = 0.5, 1, 2$ were performed. In these experiments, scale was the only source of variability between the source and target images.

The matching accuracy as a function of scale is shown in Table 2 (first row). As can be seen, the accuracy deteriorates significantly when the scale changes. Notice that the error for scale 2 is much larger than for scale 0.5; the reason is that the target image at scale 2 is much larger and therefore allows for larger errors.

TABLE 2
Accuracy of Matching as a Function of Image Scale

Image scale	0.5	1	2
Single fragment scale, constant illumination	21	1.5	67
Three fragment scales, constant illumination	10	1.5	29
Single fragment scale, varying illumination	21	6	62
Three fragment scales, varying illumination	11	6	31

The average error (in pixels) is shown for three scales. Creating candidate matches at multiple scales improves matching accuracy and, in effect, automatically determines the image scale.

4.3.2 Extended Scheme

The basic matching scheme above could not cope with scale changes because it assumed that the source and target fragments are of roughly the same size. This assumption is violated if the images are scaled with respect to each other. As described in Section 3.3, candidate target fragments of multiple sizes can be created to facilitate matching across scale.

We repeated the experiments described in Section 4.3.1 using candidate target fragments of three different scales (0.5, 1, and 2). Note that the same three candidate fragment scales were used for all image scales S. The results are shown in Table 2 (second row). One observation is that having fragments at unneeded scales does not affect the performance significantly. For example, for image scale $S = 1$, only the candidate target fragments at scale $s = 1$ are relevant, but fragments at scales $s = 0.5$ and $s = 2$ were also present and could conceivably interfere with the matching. However, Table 2 shows that this did not happen. For image scales $S \neq 1$, the performance is significantly improved relative to the naive scheme. Notice, however, that the performance does not return to the baseline level. Therefore, further improvements are desirable. One possibility is to use scale-invariant descriptors such as SIFT [22] instead of the normalized correlation of raw pixels; this is a subject for future research.

4.3.3 Matching across Both Scale and Illumination

The experiments above showed that target fragments at multiple scales improve matching across scale. However, scale was the only source of variability in those experiments. To test whether those conclusions hold when additional sources of variability are present, we performed experiments with scaled versions of the easy illumination data set. The results are reported in Table 2. The conclusion is that target fragments at multiple scales do improve performance even when additional sources of variability are present.

4.4 Buildings

Wide-baseline matching algorithms are typically evaluated on images of almost planar or piecewise-planar objects such as books or buildings. To illustrate the applicability of MBE to wide-baseline matching, we evaluate it on a data set of five buildings [32]. Several experiments were run. In each experiment, one building was picked and two images of this building were chosen as the source and the target. The remaining four buildings comprised the validation database. The usual MBE algorithm (Section 3) was applied without modifications. Examples of matches are shown in

Fig. 11. Example images. Left: Easy illumination data set. Middle: Hard illumination. Right: Cars. (See Sections 4 and 5.).

Fig. 10. The main sources of variability in this data set are the viewpoint and scale. Since buildings are piecewise planar and the transformations are relatively small, the matching task is relatively easy. Therefore, the scheme performs well even though the validation data set consisted of only four objects. This suggests that MBE is applicable to a variety of image classes and that, for easy transformations, a small validation data set is sufficient.

5 Application to Invariant Recognition

In this section, we illustrate an application of MBE to view-invariant object recognition. In the experiments, the system is presented with a single picture of an object (for example, the face of a particular individual or a specific model of a car), taken under certain viewing conditions. The task is then to identify other images of the same object in arbitrary viewing conditions (that is, other images of the same person or the same car model). The scheme that was used for recognition is an extended version of those in [6], [7], described in [19]. Briefly, an object from a general class is represented by a set of object parts, as described in Section 3.1. Each part, in turn, is represented by a set of fragments that depict this part under all relevant viewing conditions. In practice, the size of this set is not very large; for example, 15 fragments are sufficient to cover all relevant pose variations [19]. This set is called an *extended fragment*. The extended fragment is considered to be present in an image if one of its constituent fragments is present. Since each extended fragment contains information regarding the appearance of the given object part under different viewing conditions, its presence or absence in the image depends only on the object and not on the viewing conditions. Therefore, the list of extended fragments that are active in an image forms a view-invariant signature for the object and this signature is used for subsequent invariant recognition.

An important step in the scheme is the extraction of extended fragments. This step requires the matching of corresponding object parts across viewing conditions. In [19], video sequences were used to obtain the matches. Such video sequences are not applicable to illumination changes and are not always available for pose variations. The results show that the matches can be obtained by the fragment equivalence scheme described above, without using video sequences and without compromising performance.

In our experiments, four data sets were used. The pose, easy illumination, and hard illumination data sets were described in Section 4. In addition, a data set [33] consisting of 33 toy cars viewed from two widely separated directions was used (Fig. 11).

Each data set was randomly divided into training and testing groups (sizes are given in Table 3). From the training group, matching fragments were extracted using the fragment equivalence method described above. These matching fragments formed extended fragments which were used in the recognition stage to represent novel objects of the same class in an invariant manner. During recognition, a single picture of a given object (called the "target object") in certain viewing conditions was presented. The task was to recognize the target object in significantly different viewing conditions among a set of distractors. Images of objects of the same class as the target object were used as distractors. Several recognition tasks are illustrated in Fig. 12. Notice that the objects in the testing set did not appear in training and recognition therefore required generalizing from a single image of a novel object.

Some examples of extended fragments obtained automatically by the method are shown in Fig. 3. The performance of the algorithm is summarized in Table 3. As can be seen, the results are nearly perfect for the illumination tasks. The performance is somewhat reduced for the more difficult pose tasks but still remains high. The results illustrate that class-based MBE can be used successfully for the task of invariant recognition.

6 Discussion

A scheme called MBE for class-based matching of object parts was described. Appearance-based comparisons are used in this scheme only to match features in similar viewing conditions. Since comparing appearances across different viewing conditions is not needed, the scheme can perform matching even when feature appearance is significantly altered. In particular, it is not restricted to locally planar objects or piecewise-affine transformations. The scheme is applicable to a variety of natural object classes; it is completely automatic and does not require examples of correct matches.

The proposed scheme was evaluated on several different object classes and several sources of variability (such as illumination, pose, and scale). Since the scheme establishes matches between individual fragments (rather than finding a global image transformation), it can tolerate partial

TABLE 3
Percentage of Correct Recognition (Average $\pm$ St. Dev.)

Data set	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{D}\rvert$	Performance (in %)
Illumination – easy	58	10	100
Illumination – hard	58	10	98 ± 4
Pose – faces	40	10	93 ± 5
Pose – cars	28	5	88 ± 14

$\lvert\mathcal{T}\rvert$: *Number of training images,* $\lvert\mathcal{D}\rvert$: *Number of distractors.*

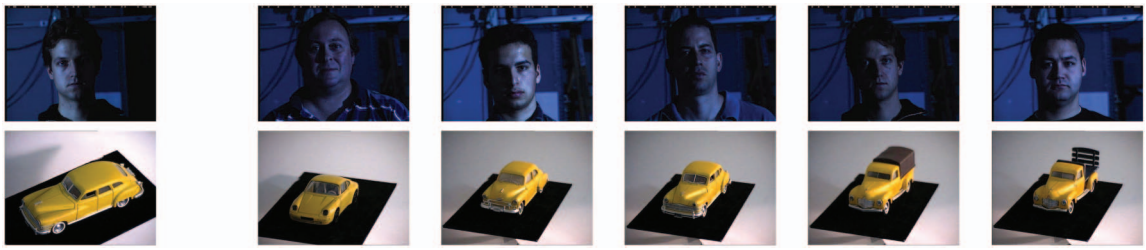


Fig. 12. Examples of recognition tasks. The single image on the far left was presented to the system for familiarization. Subsequently, the images on the right were presented and the task was to identify the object displayed on the left. Note that none of the testing images was available in training and, therefore, the identification was based on just the single image displayed on the left. Row 1: Hard illumination data set. (Only five faces are shown; the complete testing set always included 10 faces.) Row 2: Cars data set. Note the subtle difference between the second and third car images.

occlusions and, since the fragments used for matching are of significant size (10-20 percent of image size), their appearance is distinctive enough to tolerate clutter as well. This is partially confirmed by the experiments on the illumination data sets (Fig. 11), which contain some degree of background clutter. However, it is desirable to evaluate the scheme on a larger database with an even more cluttered background. This evaluation is a subject for future research.

The current implementation of the scheme used normalized correlation (NCC) to compare fragments. Although NCC is not invariant to changes of pose, scale, and nonlinear changes of illumination, the resulting scheme still gave useful performance because matching by NCC was only required within similar viewing conditions, rather than across different viewing conditions. However, in the future, it might be of interest to replace NCC with an affine-invariant measure to achieve further invariance.

MBE depends on the availability of a validation database. This is a potential disadvantage compared to generic matching approaches [2], [3], which can perform matching with just two images. The main justification for using a validation database is that it allows us to significantly improve the accuracy of matching in scenarios that are challenging for other existing approaches (Sections 1 and 2). In practice, such databases are already available for domains such as object recognition and classification.

The validation database also carries some overhead in computation. However, the object parts matched by the proposed algorithm are not object specific and are shared by approximately half of the objects in the validation database [34]. Therefore, in the process of establishing a match between two views of a single object, matches for multiple additional objects are found. The overhead of using the validation database to establish a single match is therefore compensated for by establishing multiple additional matches without additional processing. This is especially suited to tasks such as object recognition, where matches frequently need to be established for all objects in the training database.

The proposed scheme utilizes the property that, whenever two objects share a similar part, the similarity is preserved across viewing conditions. Intuitively, the similarity of appearance of two object parts in certain viewing conditions indicates the similarity of the parts themselves. Similar parts will then have similar appearance in other viewing conditions as well. However, in principle, the similarity of appearance in certain viewing conditions is not sufficient to deduce the similarity of object parts. Two parts may appear similar in some viewing conditions but not in others. Such spurious coincidences may affect fragment consistency and reduce the accuracy of matches. This issue can be addressed systematically by including in the validation database images under additional viewing conditions. The similarity of parts could then be determined based on comparisons in multiple viewing conditions. However, as the results in Section 4 indicate, this extension is not necessary to achieve reasonable performance.

Similar comments apply to the proximity constraint introduced in Section 3.2. It is well known that the proximity of two points in an image may be an artifact of viewing conditions. For example, it is quite common for two points that are far apart in one image to project to nearby locations in another image due to foreshortening. Again, this issue can be addressed systematically by learning from the validation database (including images under additional viewing conditions) which occurrences of proximity are spurious and discarding them. However, soft probabilistic constraints allow the scheme in Section 3.3 to tolerate occasional violations of the proximity constraint. The results in Section 4 indicate that high matching accuracy can be obtained without such detailed learning.

An application to the problem of invariant object recognition was presented. An additional possible application is to use class-based information in a similar manner to improve current stereo-matching techniques. Most current matching techniques for wide-baseline stereo rely on affine-invariant features. The accuracy of matches obtained by these features is reduced significantly when nonaffine transformations are present, for example, due to large pose changes of nonplanar regions, the effects of highlights, shadows, etc. For familiar objects such as faces, class-based matching techniques could be used to improve the accuracy of matches. The suggested approach is to use the fragment equivalence scheme described above to obtain matches between object parts as the first stage. Affine invariants [2], [3], [4] may then be applied locally, within small regions of the matched fragments, to refine the correspondences. Since the accuracy of the affine approximation improves for smaller regions, the matches will become more accurate. (Small regions cannot be used globally due to matching

ambiguities.) Exploring the use of the proposed framework for stereo matching and 3D reconstruction remains an interesting subject for future work.

ACKNOWLEDGMENTS

This work was supported by ISF Grant 7-0369 and IMOS Grant 3-992. Dr. Bart was a postdoctoral associate at the Institute for Mathematics and Its Applications at the University of Minnesota while performing parts of this work. Portions of the research in this paper use the FERET database of facial images collected under the FERET program [30]. An early version of this paper appeared in [35].

REFERENCES

[1] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," Technical Report CMU-CS-91-132, Carnegie Mellon Univ., Apr. 1991.
[2] T. Tuytelaars and L.V. Gool, "Wide Baseline Stereo Matching Based on Local, Affinely Invariant Regions," *Proc. 11th British Machine Vision Conf.*, pp. 412-425, 2000.
[3] K. Mikolajczyk and C. Schmid, "An Affine Invariant Interest Point Detector," *Proc. Seventh European Conf. Computer Vision*, pp. 128-142, 2002.
[4] M. Brown and D. Lowe, "Invariant Features from Interest Point Groups," *Proc. 13th British Machine Vision Conf.*, 2002.
[5] R. Basri and D. Jacobs, "Recognition Using Region Correspondences," *Int'l J. Computer Vision*, vol. 25, no. 2, pp. 141-162, 1997.
[6] S. Agarwal and D. Roth, "Learning a Sparse Representation for Object Detection," *Proc. Seventh European Conf. Computer Vision*, pp. 113-127, 2002.
[7] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual Features of Intermediate Complexity and Their Use in Classification," *Nature Neuroscience*, vol. 5, no. 7, pp. 682-687, 2002.
[8] D.I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves, "Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction," *Proc. Royal Soc. London, Series B: Biological Sciences*, vol. 223, pp. 293-317, 1985.
[9] M.J. Black and P. Anandan, "A Framework for the Robust Estimation of Optical Flow," *Proc. Fourth Int'l Conf. Computer Vision*, pp. 231-236, 1993.
[10] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721-732, July 1997.
[11] C. Wallraven and H.H. Bülthoff, "Automatic Acquisition of Exemplar-Based Representations for Recognition from Image Sequences," *Proc. CVPR 2001—Workshop Models vs. Exemplars*, 2001.
[12] S.-H. Lai, "Robust Image Matching under Partial Occlusion and Spatially Varying Illumination Change," *Computer Vision and Image Understanding*, vol. 78, pp. 84-98, 2000.
[13] G.D. Hager and P.N. Belhumeur, "Efficient Region Tracking with Parametric Models of Geometry and Illumination," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025-1039, Oct. 1998.
[14] M.J. Black and A.D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation," *Proc. Fourth European Conf. Computer Vision*, pp. 329-342, 1996.
[15] D. Tell and S. Carlsson, "Combining Appearance and Topology for Wide Baseline Matching," *Proc. Seventh European Conf. Computer Vision*, pp. 68-81, 2002.
[16] V. Ferrari, T. Tuytelaars, and L.V. Gool, "Simultaneous Object Recognition and Segmentation by Image Exploration," *Proc. Eighth European Conf. Computer Vision*, 2004.
[17] D.G. Lowe, "Three-Dimensional Object Recognition from Single Two-Dimensional Images," *Artificial Intelligence*, vol. 31, no. 3, pp. 355-395, 1987.
[18] P. Viola and W. Wells, "Alignment by Maximization of Mutual Information," *Int'l J. Computer Vision*, vol. 24, no. 2, pp. 137-154, 1997.
[19] E. Bart, E. Byvatov, and S. Ullman, "View-Invariant Recognition Using Corresponding Object Fragments," *Proc. Eighth European Conf. Computer Vision*, Part II, pp. 152-165, 2004.
[20] A. Chowdhury, R. Chellappa, and T. Keaton, "Wide Baseline Image Registration with Application to 3D Face Modeling," *IEEE Trans. Multimedia.* to appear.
[21] E. Sali and S. Ullman, "Combining Class-Specific Fragments for Object Recognition," *Proc. 10th British Machine Vision Conf.*, pp. 203-213, 1999.
[22] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[23] H. Abdi, "Kendall Rank Correlation," *Encyclopedia of Measurement and Statistics*, N.J. Salkind, ed., Sage, 2007.
[24] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete Markov Image Modeling and Inference on the Quadtree," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 390-404, 2000.
[25] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm," *IEEE Trans. Information Theory*, vol. 47, no. 2, pp. 498-519, 2001.
[26] http://vision.stanford.edu/~birch/klt/, 2007.
[27] http://www.cs.brown.edu/people/black/ignc.html, 2008.
[28] http://www.inrialpes.fr/lear/people/Mikolajczyk/, 2007.
[29] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces," Technical Report CMU-RI-TR-01-02, Robotics Inst., Carnegie Mellon Univ., Jan. 2001.
[30] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295-306, 1998.
[31] O. Chum and J. Matas, "Randomized RANSAC with $t_{d,d}$ Test," *Proc. 13th British Machine Vision Conf.*, 2002.
[32] "Oxford Colleges," http://www.robots.ox.ac.uk/~vgg/data2.html, 2008.
[33] "Weizmann Institute Toy Car Database," http://www.wisdom.weizmann.ac.il/~cars, 2008.
[34] S. Ullman and E. Bart, "Recognition Invariance Obtained by Extended and Invariant Features," *Neural Networks*, vol. 17, pp. 833-848, 2004.
[35] E. Bart and S. Ullman, "Class-Based Matching of Object Parts," *Proc. IEEE Workshop Image and Video Registration*, 2004.

Evgeniy Bart received the BSc degree in physics and computer science from Tel Aviv University in 1999 and the MSc and PhD degrees in computer science from the Weizmann Institute of Science in 2002 and 2004, respectively. He participated in the IMA annual imaging program from 2004 to 2005 and currently has a postdoctoral position at the California Institute of Technology (Caltech). His research interests include computer and human vision and machine learning.

Shimon Ullman received the undergraduate degree in mathematics, physics, and biology from the Hebrew University, Israel. He received the PhD degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT) in 1977. He was an associate and full professor at MIT until 1994 and simultaneously took a position in the Department of Computer Science and Applied Mathematics at the Weizmann Institute of Science, Israel, where he is currently the Samy and Ruth Cohn Professor of Computer Science. He is the 2008 recipient of the David E. Rumelhart prize in human cognition.